Compact representations of the articulatory-to-acoustic mapping

Blaise Potard, Yves Laprie

Compact representations of the articulatory-to-acoustic mapping

Blaise Potard, Yves Laprie

Speech Team "PAROLE", LORIA, Nancy, France

potard@loria.fr, laprie@loria.fr

Abstract

Articulatory codebooks are very often used to represent the articulatory-to-acoustic mapping. They thus need to be compact while offering a very good acoustic precision. This paper presents a method of articulatory codebook construction more general than that of Ouni [1] in the sense that the articulatory-to-acoustic mapping is approximated by multivariable polynomials. The second major contribution concerns the subdivision process which finds out the most efficient subdivision, i.e. that which minimizes the size of the codebook while guarantying a very good acoustic precision.

Experiments carried out show that the size of the codebook can be divided by a factor of 20, and simultaneously, the acoustic precision can improved by a factor of 2 by using second order polynomials together with this new construction strategy.
Index Terms: acoustic-to-articulatory inversion, codebook, polynomial interpolation.

1. Introduction

Acoustic-to-articulatory inversion[1] is a challenging problem investigated for more than 30 years now. Its goal is to recover vocal tract shapes from the speech signal for e.g. talking heads animation or articulatory feedback, for hearing-impaired people or foreign language learning.

One of the most frequent approaches is analysis by synthesis, in which an articulatory-to-acoustic model is used to compute acoustic images of articulatory vectors. However, time required to compute acoustic images may become prohibitive when many solutions are explored and therefore this mapping is represented in the form of a precomputed table, called *codebook*, made up of pairs associating an articulatory vector to its corresponding acoustic vector. In this paper, we present a new codebook construction method, substantially improving the hypercube representation proposed by Ouni[1]. The main advantage of our method is the extensive exploration of the articulatory space and the homogeneous acoustic precision achieved. This method is applied to Maeda's[2] articulatory model, but could be easily applied to different models, e.g. that of Mermelstein[3]. In the first two sections, we present the structure retained, how the mapping function is modelled, tests used to subdivide the articulatory space. Finally, we present some experiments to evaluate the frequency precision and the compactness achieved by our method.

2. Hypercuboid structure

Articulatory codebooks are intended to obtain local approximations of the articulatory-to-acoustic mapping in a fast way. This is usually done by computing acoustic images of some articulatory vectors, and by then interpolating the rest of the articulatory-to-acoustic mapping from these vectors. The mapping is thus represented by easy to invert functions in small articulatory regions around these vectors. In some cases, for example Atal[4], Larar[5] or Schroeter[6] both the acoustic and articulatory spaces are discretized and the interpolation, therefore, can be seen as a 0-th order polynomial interpolation, i.e. the mapping is constant in the articulatory regions corresponding to the sampling points. More precisely, these tables can be considered as collections of relations of the form

$$f\left(V_j \pm \Delta V_j\right) = F_j \pm \Delta F_j,$$

where f is the articulatory-to-acoustic mapping, V_j is an articulatory vector, ΔV_j a discretization step for the articulatory space, F_j an acoustic vector, and ΔF_j a discretization step for the acoustic space.

Charpentier[7], Sorokin[8] or Ouni[9] use first-order polynomial interpolation which they describe as local linear functions, or piece-wise linear functions in the case of Sorokin. Charpentier uses an interesting method, in which the articulatory space is subdivided according to the curvature of the acoustic images along specific articulatory trajectories, the points of highest curvature defining reference points, and the rest of the articulatory space being interpolated using the Jacobian matrix around these references points. Sorokin's and Ouni's methods are very similar, the main difference is the fact that Sorokin seems to determine the largest hypercuboid linear around some root articulatory vector (the acoustic linearity being tested with regard to the period of F1), whereas Ouni recursively explores the whole articulatory space and describes it as hypercubes linear around their center, the acoustic linearity being tested with regards to the first three formants frequencies. Ouni and derived methods [10, 11] are the only ones to achieve an extensive coverage of the articulatory space in their codebook with a high acoustic precision.

In some sense our method also derives from that proposed by Ouni because it builds codebooks which can yield an extensive coverage of the articulatory space, and the area where the approximation is considered valid is well defined. In our case, the elementary structure in the codebook is a hypercuboid, i.e. a multidimensional rectangle, and is therefore a simple generalization of Ouni's hypercube structure. In this work, all hypercuboids are defined around their geometric center, although there is no necessity to choose this particular point. Each elementary structure is defined by a center point, a radius vector, and a large vector representing the local approximation of the articulatory-to-acoustic function, which we describe in the next section.

Assuming the articulatory space considered is N-dimensional, the mathematical definition of a hypercuboid as

[1]This work is part of the ASPI project funded by the IST Program of the Commission of the European Communities as project number IST-2005-021324.

used in our method can be expressed as:

$$H_c = \{P_0 + x, x \in \mathbb{R}^N \, | \forall i \in \{1..N\} \, |x_i| \leq r_i\},$$

where P_0 is the (geometric) center of the hypercuboid, and the N-dimensional vector r is the radius of the hypercuboid. This definition is voluntarily a restriction of the most general definition of a hypercuboid, in which all hypercuboids have the same orientation as the canonical base of the N-dimensional articulatory space. In this way, all mathematical definitions are simpler, and we can achieve a tessellation of the articulatory space.

3. Modeling the articulatory-to-acoustic relation

One of the most crucial parts when approximating the articulatory-to-acoustic (from now on called a-to-a) mapping is the choice of the modeling function. In theory, any approximation function which is simple enough to invert could be used to investigate articulatory-to-acoustic inversion, but almost all existing codebooks methods only use polynomial approximations of degrees 0 or 1.

This paper presents a more general method in which any polynomial approximation is possible. It is interesting to investigate degrees higher than one in order to explore the acoustical behavior of the articulatory model. The utility of higher degree approximation for inversion purposes is more questionable, since factorizing multivariable polynomials of degrees higher than 1 is a fairly difficult task. However, it can still be achieved, in a not very satisfying way, by locally linearizing such polynomials when needed.

3.1. Multi-variables polynomials

As the reader may not be familiar with multi-variable polynomials algebra, we present a short description of the use of such objects.

A *polynomial* of variables $x_1, \cdots, x_N$ in a ring R, denoted as $P(x_1, \cdots, x_N)$, is a sum of monomials. A *monomial* is a term of the form:

$$c.{x_1}^{k_1} {x_2}^{k_2} \cdots {x_N}^{k_N},$$

where $c \in R$ is the coefficient of the monomial, and $k_i \in \mathbb{N}$ is the exponent of variable x_i. The *degree* of a monomial is defined as $\sum_{i=1}^{N} k_i$. The degree of a polynomial is as usual defined as the maximum of the degrees of all its monomials.

We define X as polynomial $(x_1, x_2, \cdots, x_N)$, which corresponds to the polynomial function: $x_1 + x_2 + \cdots + x_N$. X^n is thus equal to: $(x_1 + x_2 + \cdots + x_N)^n$. We have the following relation:

$$X^n = \sum_{1 \leq i_1, i_2, \cdots, i_n \leq N} x_{i_1} x_{i_2} \cdots x_{i_n}$$

If R is a commutative ring (this will always be the case for us, since we are only working on $\mathbb{R}$ or $\mathbb{C}$), many of the terms of this sum can be grouped, and it can be rewritten using the multinomial formula.

To simplify notations, polynomials $P(X)$[2] (of degree n) will be noted as such:

$$P(X) = A_0 + A_1 X + A_2 X^2 + \cdots + A_n X^n.$$

[2]It should be noted that in this paper, depending on the context, X should be seen as a variable of the product ring R^N, as the polynomial $P(X) = X$, or as the polynomial function $R^N \mapsto R : (x_1, \cdots, x_N) \rightarrow x_1 + \cdots + x_N$.

In this expression, A_0 is an element of R, $A_1 = (A_{1,1}, \ldots, A_{1,N}) \in R^N$ is the coefficient vector for X, i.e.

$$A_1 X = A_{1,1} x_1 + A_{1,2} x_2 + \cdots + A_{1,n} x_n.$$

Likewise, $\forall m \in \{1..n\}, A_m \in R^{\binom{N+m-1}{m}}$ is the coefficient vector for X^m, with

$$A_m X^m = \sum_{\sum_{i=1}^{N} k_i = n} A_{m,w} \binom{n}{k_1, k_2, \ldots, k_N} x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N},$$

where w is an index for the vector $k1, k2, ..., k_N$.

3.2. Using multi-variable polynomials approximations in codebook construction

We denote as P^N the space of articulatory vectors, and A^M the space of acoustic vector. In our case, P^N is the 7-dimensional space of parameters for Maeda's articulatory model, and A^M is the space of the first three formant frequencies. We note the articulatory-to-acoustic function as $F : P^N \mapsto A^M$, and $F_i : P^N \mapsto A, 1 \leq i \leq M$ its restriction to each component of the acoustic vector.

Let us consider an elementary structure of the codebook, in our case a hypercuboid of the space of articulatory parameters. For each component i of the acoustic vector, we wish to find the polynomial $P_i(X)$ (of degree n) which best describes $F_i : P^N \mapsto A$ in this structure. This can be done simply by solving a system of equations of the form $\{F_i(X_j) = P_i(X_j)\}$ for many articulatory vectors $X_j = (x_{j1}, x_{j2}, \ldots, x_{jN})$. For that purpose, many articulatory points, at least as many as the number of the coefficients to determine, are chosen within the elementary structure; their acoustic images are calculated through the acoustic simulation. Then, after rewriting the relations $F_i(X_j) = P_i(X_j)$ as linear equations of the unknown coefficients, the over-determined set of equations is solved by using Singular Value Decomposition.

An alternative to this method would be the computation of Taylor's formula around a particular point in the hypercuboid. However, relevant derivatives needed by a Taylor development cannot be computed easily, and this development is only valid in the vicinity of a given point.

Let C_i be the vector composed of all the coefficients:

$$C_i = (\underbrace{A_0}_{1} | \underbrace{A_1}_{N} | \cdots | \underbrace{A_n}_{\binom{N+n-1}{n}}).$$

Let W_j be the vector composed of the X_j^k vectors:

$$W_j = (\underbrace{{X_j}^0}_{1} | \underbrace{{X_j}^1}_{N} | \cdots | \underbrace{{X_j}^n}_{\binom{N+n-1}{n}}) =$$

$$\left(1 \mid x_{j1}, \ldots, x_{jN} | \cdots | {x_{j1}}^n, \ldots, \binom{n}{k_1, k_2, \ldots, k_N} {x_{j1}}^{k_1} {x_{j2}}^{k_2} \cdots {x_{jN}}^{k_N}, \ldots, {x_{jN}}^n\right)$$

The equation $F_i(X_j) = P(X_j)$ can thus be rewritten as the linear equation $F_i(X_j) = W_j \, C_i$. Finally, let denote B_i as the vector containing all $F_i(X_j)$, and W as the matrix containing all W_j:

We obtain the system of linear equations:

$$W\ C_i = B_i \quad (1)$$

For this system to be over-determined, the number m of articulatory vectors to sample must be at least equal to the number of distinct coefficients, i.e.:

$$m \geq \sum_{k=0}^{n} \binom{N+k-1}{k} = \binom{N+n}{n}$$

Solving this system using SVD enables us to find the coefficients of the polynomial that minimizes the least-squares approximation error. Repeating this process for each acoustic component F_i thus gives us an optimal (in the sense of the least-squares error) polynomial approximation of degree n for the local a-to-a mapping.

One crucial part is the choice of the points used for sampling the a-to-a mapping in the hypercuboid. Unfortunately, we do not have enough space to discuss this point in detail. For the interpolation of degrees lower than or equal to three, our method currently uses the *vertices* (i.e. the corners) of the hypercuboid, the middles of all segments linking two vertices, and several points around the geometric center. These points are used to ensure smooth transitions between hypercuboids. For higher degrees, we sample some additional random points. Some samplings are more appropriate for the purpose of polynomial approximation, in particular Chebyshev sampling, and the design of a more efficient sampling scheme will be addressed in a future work.

4. Exploring the articulatory space

The polynomial approximation method allows us to get a compact representation of the approximation of the local a-to-a mapping in a small structure. In this section we present how the whole articulatory space can be divided in such structures. The method presented here is a variation of Ouni[1], and achieves much better results.

4.1. Baseline algorithm

The basic principle of Ouni's method is a recursive subdivision of the articulatory space in small structures until the local a-to-a mapping is smooth enough to respect a specified acoustic precision threshold, or the structure is too small. Structures used by Ouni are hypercubes, and when the acoustic precision in an hypercube of side r is not sufficient, this root structure is subdivided in all the sub-hypercubes of side $r/2$ it contains, and then each sub-hypercube is recursively checked. If either the center point or any of the 2^N vertices of an hypercube is in the "forbidden space" (i.e. with a very narrow constriction or a complete closure), the hypercube is systematically subdivided (or rejected if it becomes too small). In a N-dimensional articulatory space, each additional level of subdivision implies the recursive exploration of 2^N sub-hypercubes, i.e. 128 for Maeda's articulatory model since there are 7 articulatory parameters, although several of them have probably almost the same acoustic behavior, and could be regrouped in larger structures.

This paper addresses this point by introducing the hypercuboid structure, which is much more flexible than the hypercube. In our case, when the acoustic test fails in a hypercuboid of radius r, it is also subdivided, but for only one component $j \in \{1..N\}$ of the articulatory space at a time. This means that sub-hypercuboids have a radius r' defined as such: $\forall\, k \neq j, r'_k = r_k; r'_j = r_j/2$. The difficulty is then to choose the "right" direction, i.e. the one that will minimize the number of subdivisions. Next subsections address the nature of the acoustic test, and the methods used to choose the best direction for subdividing the hypercuboids.

4.2. Acoustic tests

The acoustic test simply consists in comparing acoustic images obtained through two different methods: using the articulatory synthesizer, and using the polynomial approximation. If the acoustic distance between both images is beyond a given acoustic threshold, then the test fails, and the hypercuboid will be either subdivided, or rejected if it is too small. The acoustic distance computed is usually expressed in the Bark perceptive scale, and is the maximum distance over all the points computed in the hypercuboid (and not the average distance). Indeed, the maximum distance is more meaningful to measure the acoustic irregularities than the average acoustic distance, which is always very low: indeed, when disabling the acoustic test (i.e. subdivision only occurs when part of the structure is in the forbidden space), even with first order approximation, we achieve an average resynthesized distance below 30Hz for all 3 formants frequencies.

The formula used for the acoustic test is thus :

$$\text{Test} = \max_{X \in Hc^*} (d_{\text{ac}}(F(X), P(X))) < \text{Ac\_threshold},$$

where, for an articulatory vector X, $d_{\text{ac}}(F(X), P(X))$ denotes the acoustic distance between the actual acoustic image obtained using the articulatory synthesizer $F(X)$, and the acoustic image obtained using the local polynomial approximation $P(X)$. d_{ac} in our case is the maximal distance over the first three formants frequencies, expressed in Bark. Ac_threshold denotes the acoustic threshold (usually 1 Bark in our case). Hc^* is a subset of points of the hypercuboid, usually the points already synthesized in the hypercuboid (i.e. the vertices, the middles points, and the points around the center used to compute Taylor's approximation).

4.3. Direction of subdivision

The most crucial point of this hypercuboid method is to find the "right" direction for subdivision, i.e. the direction which will yield the fewest subdivisions. The algorithm to find this direction has to be fast and accurate. For now, a simple heuristic is being used, which tries to find the direction in which the irregularity appears to be the strongest: for all directions, i.e. for all the articulatory parameters, a score is computed from the acoustic distance between synthesized vectors and the approximation obtained using Taylor's approximation in the center. Indeed, it is more meaningful to use Taylor's approximation and not the optimal polynomial in that case, because the optimal polynomial smoothes the irregularities (and therefore the error becomes almost homogeneous in all directions). The precise formula used is the following:

$$\forall i \in \{1..N\}, s_i = \frac{1}{|Hc^*|} \sum_{X \in Hc^*} \frac{|x_i|}{|X|} * d_{\text{ac}}(F(X), P_t(X)),$$

where P_t is Taylor's polynomial at the center. The direction chosen is the one that maximizes s_i. This simple heuristic is compared to the hypercubic subdivision (all directions at a time), and to a very basic scheme that always subdivides the largest component of the radius (choosing the component with the smallest index in case of equalities).

Table 1: *Summary of codebook experiments. Column "n" is an identifier for the codebook, column "d." corresponds to the degree of the polynomial approximation, "Ac" corresponds to the acoustic threshold (in Bark), "S" to the subdivision scheme (0 is the hypercubic subdivision, 1 is the basic subdivision, 2 is the subdivision in the direction of maximal perturbation, "# Hc." is the number of hypercuboids in the codebook, "vol." is the hypervolume of the codebook, and the "ΔFi" correspond to the average absolute error of resynthesis (in Hz) for respectively F1, F2 and F3.*

n	d.	Ac.	S.	#Hc.	vol.	ΔF1	ΔF2	ΔF3
1	1	$+\infty$	0	3043	779.1	10.0	15.1	14.7
2			1	643	779.1	11.7	18.1	17.9
3			2	140	782.1	13.1	19.7	21.1
4		1	0	5835	778.9	4.8	6.9	6.6
5			1	1413	778.9	7.0	9.9	8.8
6			2	464	778.9	6.8	11.1	13.2
7		0.3	1	4080	778.9	4.8	6.8	6.9
8			2	3840	778.9	4.9	7.1	7.2
9	2	$+\infty$	0	3043	779.2	4.7	6.2	5.8
10			1	643	779.1	5.4	7.5	7.8
11			2	140	782.1	5.0	8.0	9.8
12		1	1	671	778.9	5.4	7.6	7.9
13			2	164	782.1	**4.6**	**7.9**	**9.7**
15		0.3	2	750	778.9	2.5	3.4	4.8
16	3	1	2	142	782.1	1.6	3.2	5.7
17		0.3	2	378	779.1	1.2	2.0	3.5
18	4	1	2	140	782.1	0.6	1.4	2.7
19		0.3	2	271	779.1	0.5	1.1	2.2

5. Codebook experiments

Experiments were conducted on a subpart of the articulatory space ($[0, 3]$ for the 7 articulatory parameters) by creating a series of codebooks. To limit subdivisions, hypercuboids have a minimal hypervolume fixed at 0.1.

Several parameters were varied for the construction of the codebooks: degree of interpolation (1 to 4), acoustic threshold ($+\infty$, 1 Bark, 0.3 Bark), and scheme of subdivision (hypercubic, larger side first, maximum perturbation). For degree 3 and 4, only maximum perturbation scheme is presented, and only for acoustic thresholds 1 and 0.3 Barks. For each codebook, we computed the number of hypercuboids needed to achieve the given acoustic precision, the volume of the articulatory space covered, the average and maximum error for resynthesis of the three formants frequencies. In each codebook, the average acoustic precision was evaluated by generating 10000 random points within the articulatory space, and by generating and comparing the acoustic images obtained either by finding the corresponding hypercuboid and using the local polynomial interpolation, or using the articulatory synthesizer.

Table 1 summarizes the codebooks experiments. The $+\infty$ acoustic threshold was used to find out what was the minimal fragmentation we could hope to achieve. It allows us to see that even when disabling the acoustic test, the hypercubic subdivision fragments a lot (at least in this particular part of the articulatory space), since the average volume of an hypercube in the codebook 1 is about vol / #Hc $= 0.25$ for a minimal volume of 0.1. We can see that directed subdivision schemes 1 and 2 are a lot better, since their average volumes are respectively 1.2 and 5.6. Not surprisingly, for degree 1 and 2 there are exactly the same number of hypercuboids in this case. Interestingly enough, we observe that the average errors are always below 30Hz, even with no acoustic test.

We observe that by increasing the degree of the polynomials, we achieve better acoustic precision. With degree 4, the average error is below 3Hz for all formants. Scheme 2 allows a codekook size about 20 times smaller than the hypercubic division, for a small acoustic degradation (about 20% less precise than hypercubic division). Codebook 13 (degree 2, with subdivision scheme 2) appears to be the best compromise in term of size of the codebook and acoustic precision: about the same acoustic precision as codebook 4, but in a much more compact form (second order polynomial hypercuboids take about four times more space on disk as first order polynomials hypercuboids, therefore we gain almost a factor 10 on codebook size).

6. Concluding remarks

This study demonstrates a new articulatory codebook construction method that allows very precise resynthesis of acoustic vectors for Maeda's articulatory model, and a very compact representation of the articulatory-to-acoustic mapping. Further work can still be done to improve this method: using different sampling points to determine the best-fit polynomials and investigate the use of different acoustic vectors (e.g. cepstral coefficients).

7. References

[1] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[2] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires," in *Actes 10èmes Journées d'Etude sur la Parole*, Grenoble, Mai 1979, pp. 152–162.

[3] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, pp. 1070–1082, 1973.

[4] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[5] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE Trans. ASSP*, vol. 36, no. 12, pp. 1812–1818, December 1988.

[6] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, Part. II, pp. 133–150, January 1994.

[7] F. Charpentier, "Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic non-linearities," *Speech Communication*, vol. 3, pp. 291–308, 1984.

[8] V. Sorokin and A. Trushkin, "Articulatory-to-acoustic mapping for inverse problem," *Speech Communication*, vol. 19, pp. 105–118, 1996.

[9] S. Ouni and Y. Laprie, "Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire," in *Actes des Journes d'tude sur la parole, Aussois*, Jun. 2000.

[10] B. Potard, Y. Laprie, and S. Ouni, "Expériences d'inversion basées sur un modèle articulatoire," in *Journées d'Etudes sur la Parole - JEP'04 , Fs, Maroc*, Apr 2004. [Online]. Available: http://www.loria.fr/publications/2004/A04-R-335/A04-R-335.ps

[11] B. Potard and Y. Laprie, "Using phonetic constraints in acoustic-to-aticulatory inversion," in *Interspeech, Lisboa*, Sep. 2005, pp. 3217–3220.